Title
A Unified Knowledge Representation System for Robot Learning and Dialogue

Author
Shukla, Nishant

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Unified Knowledge Representation System for Robot Learning and Dialogue

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Computer Science

by

Nishant Shukla

2016

ABSTRACT OF THE THESIS

A Unified Knowledge Representation System

for Robot Learning and Dialogue

by

Nishant Shukla

Master of Science in Computer Science

University of California, Los Angeles, 2016

Professor Song-Chun Zhu, Chair

To allow wide-spread adoption of consumer robotics, robots must be able to adapt to their environment by learning new skills and communicating with humans. Each chapter explains a contribution to achieve this goal. Chapter One covers a stochastic And-Or knowledge representation framework for robotic manipulations. Chapter Two further expands this established system for robustly learning from perception. Chapter Three unifies perception with natural language for a joint real-time processing of information. We've successfully tested the generalizability and faithfulness of our robotic knowledge acquisition and inference pipeline. We present proof of concepts in each of the three chapters.

The thesis of Nishant Shukla is approved.

Stefano Soatto

Yingnian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2016

Table of Contents

List of Figures

Acknowledgements

Chapter One is a reprint of a short-paper I published to the 2015 AAAI Fall Symposium on AI for Human-Robot Interaction (AI-HRI 2015). The title of the work is “A Unified Framework for Human-Robot Knowledge Transfer.” The list of co-authors is N Shukla, C Xiong, and SC Zhu. I was involved in developing the perception system and writing the paper, Caiming Xiong developed the system architecture, and Professor Song-Chun Zhu directed the study.

Chapter Two is a reprint of a conference paper published to the International Conference on Robotics and Automation (ICRA 2016). The title of the work is “Robot learning with a spatial, temporal, and causal and-or graph.” The list of co-authors is C Xiong, N Shukla, W Xiong, SC Zhu. Caiming and I contributed equally to the work, developing the perception learning system and writing the paper. Wenlong Xiong helped develop the simulation software. Professor Song-Chun Zhu directed the study.

Chapter Three is a reprint of a workshop paper published to the AAAI'16 Workshop on Symbiotic Cognitive Systems. The title of the work is “Task Learning through Visual Demonstration and Situated Dialogue.” The list of co-authors is C Liu, JY Chai, N Shukla, SC Zhu. Chaongsong Liu and Joyce Chai integrated the language component. Professor Zhu and I were in charge of the vision and casual learning component.

This work was supported by a DARPA SIMPLEX grant N66001-15-C-4035, DARPA MSEE project FA 8650-11-1-7149. In addition, we would like to thank SRI International and OSRF for their support.

Chapter 1: A Unified Framework for Human-Robot Knowledge Transfer

Abstract

Robots capable of growing knowledge and learning new tasks is of demanding interest. We formalize knowledge transfer in human-robot interactions, and establish a testing framework for it. As a proof of concept, we implement a robot system that not only learns in real-time from human demonstrations, but also transfers this knowledge.

Introduction

Transferring knowledge is a vital skill between humans for efficiently learning a new concept. In a perfect system, a human demonstrator can teach a robot a new task by using natural language and physical gestures. The robot would gradually accumulate and refine its spatial, temporal, and causal understanding of the world. The knowledge can then be transferred back to another human, or further to another robot. The implications of effective human to robot knowledge transfer include the compelling opportunity of a robot acting as the teacher, guiding humans in new tasks.

The technical difficulty in achieving a robot implementation of this caliber involves both an expressive knowledge structure and a real-time system for non-stop learning and inference. Recently, skill acquisition and representation have become some of the core challenges in achieving robots capable of learning through human demonstration.

We propose a real-time unified learning and inference framework for knowledge acquisition, representation, and transfer. Knowledge is represented in a Spatial, Temporal, and Causal And-Or Graph (STC-AoG) hierarchical network (Tu et al. 2014), which can be thought of as a stochastic grammar. The STC-AoG encapsulates the hierarchical compositional structures of physical objects, logical deductions, and instance-based actions. Our knowledge graph manipulation framework enables learning to be a continuous on-line process that occurs alongside inference. We view a robot as a knowledge database, where humans may deposit and withdraw skills. These skills can be used by both humans and robots alike.

As a proof of concept, we teach an industrial robot how to fold clothes (Figure 1). The robot watches a human demonstrator and learns in real-time. To test the faithfulness of the human-robot knowledge transfer, we propose an evaluation procedure called the Knowledge Transfer Test. Our experiments demonstrate that our proposed framework can adequately transfer knowledge to and from a robot. Furthermore, we illustrate our system's interactive learning capabilities that are backed by a Bayesian formulation.

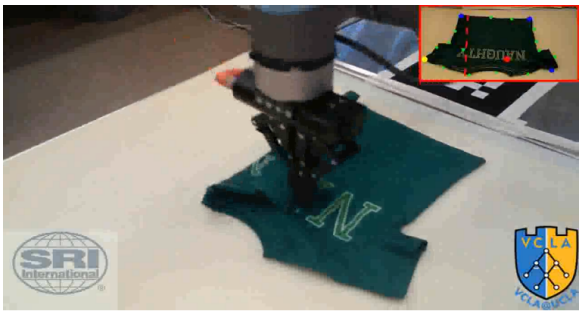


Figure 1: The robot autonomously performs a cloth folding task after learning from a human demonstration.

Related Works

We extend the learning of And-Or grammars and semantics from video (Si et al. 2011) to an interactive real-time robotics platform with a natural communication interface between humans. The And-Or data structure has also been used in learning a visually grounded storyline model from labeled videos (Gupta et al. 2009); however, our system requires no labeled data, and evokes a richer segmentation of spatial, temporal, and causal concepts for more tractable queries. Miller et al. establish high standards for a cloth-folding robot, but our focus is instead on novel learning, knowledge representation, and knowledge transfer. The action-planning inference system in our STC-AoG data structure resembles closest to a Planning Graph (Blum and Furst 1997),

which is essentially an STC-AoG without causal nodes. Yang et al. learn concrete action commands from small video clips. Unlike their system, our design allows a modifiable grammar and our performance is measured on multi-step actions.

Contributions

The contributions of our paper include the following:

- A unified real-time framework for learning and inference on a robot system by using an STC-AoG data structure for ensuring faithful knowledge transfer.
- A test to evaluate the success of a human-robot knowledge transfer system.

Our Approach

We encapsulate knowledge by an expressive graphical data structure $G_\Omega = (G_s, G_t, G_c)$ which models the compositional structures of objects G_s, actions G_t and causality G_c. A specific piece of information or skill, such as how to fold clothes, is a subgraph $G \subseteq G_\Omega$. The goal of knowledge transfer is to deliver G from one agent (e.g. human) to another (e.g. robot) with a minimum loss of knowledge.

In human-robot interactions, we restrict communication to only physical actions through a video-camera sensor V, and natural language text L. Therefore, the learner must construct an optimal G based only on V and L, resulting in the Bayesian formulation,

$$G^* = \arg\max_{G_t} P(G_t|V,L) = \arg\max_{G_t} \frac{P(V|G_t,L)P(G_t,L)}{P(V,L)}$$

Similar to (Ha, Kim, and Zhang 2015), we use a graph Monte Carlo method that assumes the graph structure is determined only by that of the previous iteration.

$$G^* = \arg\max_{G_t} P(V|G_t,L)P(G_{t-1},L)$$

The learning algorithm is similar to a marginalization-based parameter learning algorithm, where we first marginalize our STC-AoG, and learn the S-AoG, T-AoG and C-AoG separately, then jointly learn the conditional model between each other.

Figure 2 shows a small segment of G^*, and specific details of the spatial, temporal, and causal segments are described as follows.

Spatial Representation

Sensory data from the environment is encoded to form a belief representation. We use a PrimeSense camera to capture RGB-D (Red, Green, Blue, and Depth) information per frame. We represent every cloth by a high-level abstract understanding based off its contour shape, and a low-level representation by specific keypoints. The keypoints and contour shape data are used as input to the folding algorithm which generalizes to arbitrary articles of clothing. To store the hierarchical structure of physical objects, we use an And-Or

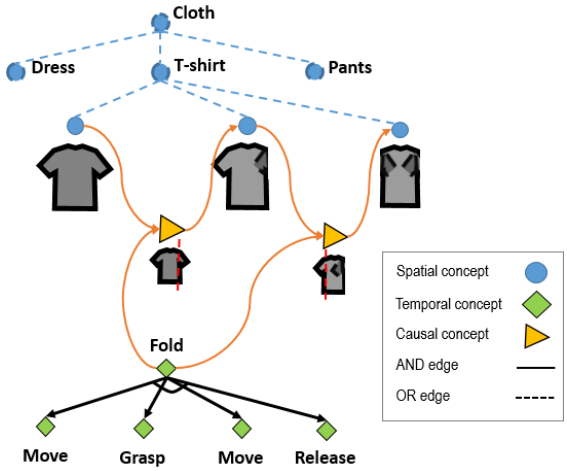


Figure 2: An automatically learned STC-AoG.

Graph data-structure, called the Spatial And-Or Graph (S-AoG) (Zhu and Mumford 2006). **AND** nodes in the S-AoG represent structural compositionality (i.e. a vehicle has an engine). **OR** nodes in the S-AoG represent variations (i.e. a car is a type of vehicle).

Causal Representation

The perceived model of the world is then used to learn a logical cause-and-effect type of reasoning from a single-instance, inspired by the Dynamics Model (Wolff 2007).

The Dynamics Model defines causal relationships as interpretations of force vectors. The nodes in the S-AoG are normalized feature vectors in a higher dimensional space, and are acted on by force vectors from the T-AoG. As per the model, if the net force on a spatial node is collinear with the vector represented by the end-state of an action, then a causality is deduced, as shown in Figure 3.

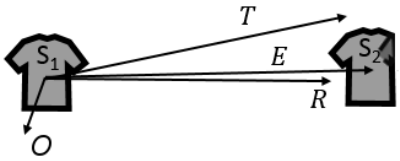


Figure 3: Forces associated with the simulated actuator T and other forces O sum to produce a resulting R force, which is nearly collinear to the observed end-state E, implying that T causes S_1 to become S_2.

The causal relationships are stored in a Causal And-Or Graph (C-AoG). **AND** nodes in the C-AoG indicate that all preconditions are necessary, whereas **OR** nodes indicate that only one of the preconditions is sufficient.

Temporal Representation

These deductive models are used to plan out the next course of action, which may affect the environment. The actuators that affect the environment, whether by the robot or the human, are represented in another data-structure, called the Temporal And-Or Graph (T-AoG). **AND** nodes represent actions done in a series of steps. **OR** nodes represent variations in possible actions.

Joint Representation

We represent the physical models (S-AoG), the reasoning models (C-AoG), and the environment actuators (T-AoG) all into one unified Spatial Temporal Causal And-Or Graph (STC-AoG) data structure. As a consequence, the whole system forms a closed-loop from perception to learning to inference, and back again to perception. Figure 2 demonstrates a small portion of the STC-AoG applied to a cloth-folding task.

Knowledge Transfer Test

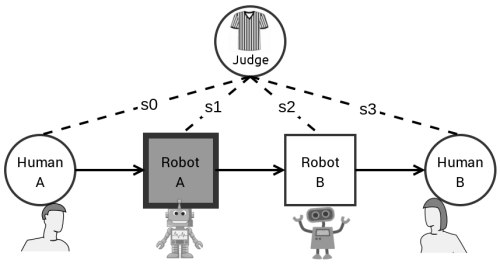


Figure 4: Arrows represent the direction of knowledge transfer. The judge assigns task scores at each step.

One of the most useful properties of knowledge transfer is the ability to propagate the knowledge among others. To determine the proficiency of knowledge transfer to and from an artificial agent, we propose the following three-part test.

A human demonstrator H_A will perform a chosen task to receive a task score s_0 by a human judge. In the first part of the test, H_A will teach this task to a robot R_A that has not been previously trained on the task. The judge will assign a task score s_1 based on R_A's performance.

Next, the second test will evaluate R_A's ability to transfer knowledge to another robot R_B that has not been previously trained on the task. Robot-to-robot knowledge transfer can be as direct as sending over the explicit knowledge structure, which in our case is the STC-AoG. Again, the judge will assign a task score s_2.

Finally, R_B must teach the task to a human H_B that has no previous knowledge of the task procedure. A task score s_3 will be assigned by the judge. If all three task scores match within 10% of s_0, then R_A is said to have passed the Knowledge Transfer Test (KTT). The entire process is visualized in Figure 4

Experimental Results

We evaluate our framework on a two-armed robot using the proposed Knowledge Transfer Test on a cloth folding task. To benchmark real-time performance, we calculate the ratio between the duration of the demonstration and the total time spent learning. The average speed of our robot system is 5 fps, resulting in a system which out-performs most perception-heavy robot learning-systems today.

Our robot was able to understand the cloth-folding task, generating a STC-AoG similar to Figure 2, confidently enough to pass the first part of the KTT. We were able to save the graphical structure and load it into a different type of robot to pass the second part of the KTT. The robot was also able to teach the task successfully to a human, but since folding clothes is already a well known skill by most humans, we set aside deeper investigation of robot-to-human teaching for future work.

Acknowledgments

This work is supported by the Office of Naval Research grant N000141010933 and the DARPA Award N66001-15-C-4035. In addition, we would like to thank SRI International and OSRF for their hardware support.

References

Blum, A. L., and Furst, M. L. 1997. Fast planning through planning graph analysis. *Artificial Intelligence* 90(1):281–300.

Gupta, A.; Srinivasan, P.; Shi, J.; and Davis, L. S. 2009. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2012–2019. IEEE Computer Society.

Ha, J.-W.; Kim, K.-M.; and Zhang, B.-T. 2015. Automated construction of visual-linguistic knowledge via concept learning from cartoon videos. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015), Austin*.

Miller, S.; van den Berg, J.; Fritz, M.; Darrell, T.; Goldberg, K.; and Abbeel, P. 2012. A geometric approach to robotic laundry folding. *International Journal of Robotics Research (IJRR)* 31(2):249–267.

Si, Z.; Pei, M.; Yao, B.; and Zhu, S.-C. 2011. Unsupervised learning of event and-or grammar and semantics from video. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 41–48. IEEE.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S.-C. 2014. Joint video and text parsing for understanding events and answering queries. *MultiMedia, IEEE* 21(2):42–70.

Wolff, P. 2007. Representing causation. *Journal of experimental psychology: General* 136(1):82.

Yang, Y.; Li, Y.; Fermuller, C.; and Aloimonos, Y. 2015. Robot learning manipulation action plans by unconstrained videos from the world wide web. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Zhu, S.-C., and Mumford, D. 2006. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2(4):259–362.

Chapter 2:
Robot Learning with a Spatial, Temporal, and Causal And-Or Graph

Abstract— **We propose a stochastic graph-based framework for a robot to understand tasks from human demonstrations and perform them with feedback control. It unifies both knowledge representation and action planning in the same hierarchical data structure, allowing a robot to expand its spatial, temporal, and causal knowledge at varying levels of abstraction. The learning system can watch human demonstrations, generalize learned concepts, and perform tasks in new environments, across different robotic platforms. We show the success of our system by having a robot perform a cloth-folding task after watching few human demonstrations. The robot can accurately reproduce the learned skill, as well as generalize the task to other articles of clothing.**

I. INTRODUCTION

Writing automated software on robots is not nearly as robust as that on traditional computers. This is due to the heavy burden of matching software assumptions to physical reality. The complexities and surprises of the real world require robots to adapt to new environments and learn new skills to remain useful.

In robot automation, implicit motor control is widely used for learning from human demonstrations [1] [2] [3]. However, implicit motor control is insufficient for generalizing robot execution. For instance, a robot can imitate a human's demonstration to open a door; yet, it cannot execute a similar motion trajectory such as opening a window without the explicit representation of the task. Intuition such as how to rotate the joints of an arm is not something easily expressible, but rather learned through experiences. Uniting explicit and implicit knowledge allows immediate communication through natural language [8], as well as clear grounding of abstract concepts into atomic actions.

In this paper, we propose a unified framework to bridge the implicit motor control with explicit high-level knowledge so the robot can understand human behavior, perform a task with feedback control, and reason in vastly different environments. As a proof of concept, we teach a robot how to fold a shirt through few human demonstrations, and have it infer how to fold never-before-seen articles of clothing, such as pants or towels. The same causality-learning framework can be extrapolated to arbitrary tasks, not just cloth-folding. Specifically, the robot can learn different skills (e.g. flattening, stretching) depending on which features it tracks (e.g. smoothness, elastic stress). Moreover, since explicit knowledge is structured graphically, our framework naturally allows for the merging, trimming, and addition of knowledge from various human demonstrations, all with feedback control. The high-level concepts are human-understandable, so both the human and robot can communicate through this intermediate language [7]. Thus, programming the robot becomes an act of merely modifying a graph-based data structure.

The contributions of this paper include the following:

- Proposes a cross-platform stochastic framework for robots to ground human demonstrations into hierarchical spatial, temporal, and causal knowledge.
- Demonstrates a robot capable of learning, correcting its mistakes, and generalizing in a cloth-folding task from human demonstrations.
- Establishes the first system to use a non-rigid physical simulation to model the robot's environment to improve task execution.
- Provides experimental evidence of our framework to generalize a cloth-folding task across different clothes and different robot platforms.

II. RELATED WORKS

While precisely grounding a human demonstration to atomic robot actions has been done in various forms [6] [13] [26], we instead focus on the novel representation and generalizability of tasks. Beetz et al. integrate robot knowledge representation into the perception processes as well, but our framework allows alternative planning generated by probabilistic sampling to match observed expectations. For example, there are multiple ways to fold a t-shirt, and each of these ways has its own likelihood. Our probabilistic learning framework resembles closest to the human-inspired Bayesian model of imitation by Rao et la. [21]. However, we instead

C. Xiong, N. Shukla, W. Xiong, and S.-C. Zhu are with the Center for Vision, Cognition, Learning, and Autonomy (VCLA), University of California, Los Angeles caimingxiong@ucla.edu, nxs@ucla.edu, wenlongx@gmail.com, sczhu@stat.ucla.edu

emphasize the hierarchical and ever-changing nature of spatial, temporal, and causal concepts in the real world.

Autonomously folding clothes has been demonstrated in various works. Wang et al. [29] were able to successfully design a perception-based system to manipulate socks for laundry. Miller et al. [11] have demonstrated sophisticated cloth-folding robots, and Doumanoglou et al. [28] have made substantial progress in autonomously unfolding clothes. On the other hand, our focus is to understand how to perform arbitrary tasks. There are other systems [6] that also learn concrete action commands from small video clips, but unlike those, our design allows a modifiable grammar and our performance is measured on multi-step long-term actions. Furthermore, our solution to knowledge representation is more powerful than commonsense reasoning employed by first-order logic [19], since it takes advantage of the probabilistic models under ambiguous real-world perception.

Our work is based on the knowledge representation system incorporated by Tu et al. [12], augmented heavily into the robotics domain. We extend the learning of event And-Or grammars and semantics from video [4] to our real-time robotics framework. The And-Or graph encapsulates a conformant plan under partial observability, enabling an architecture that is cognitively penetrable since an updated belief of the world alters the robot's behavior [14]. Unlike traditional graph planning [10], the hierarchical nature of the knowledge representation system enables a practical way of generating actions for a long-term goal.

III. METHOD

There is often a fine distinction between memorization and understanding, where the latter enables generalizing learned concepts. In order to understand a human task from demonstrations/videos such as cloth-folding, a knowledge representation system is necessary to ensure actions are not simply memorized. Four types of knowledge are important for understanding and generalizing:

- **Spatial knowledge** expresses the physical configuration of the environment when performing the task. For a cloth-folding task, a table, cloth, and each part of the cloth, such as the left and right sleeve of a shirt, needs to be detected.
- **Temporal knowledge** reveals the series of human actions in the process of the task. In cloth-folding, the hand motion, grip opening, and grip closing actions are essential. These actions combine together to form a fold action.
- **Causal knowledge** conveys the status change of an object in each dynamic human action. For example, a shirt may be folded in various ways, either by folding the left sleeve into the middle and then the right sleeve, or vice versa. Folding a cloth requires multiple hierarchical steps for reasoning.
- **The interplay between the spatial, temporal, and causal concepts** manifests a generalizable form of knowledge to be used in changing application domains. The robot must choose an action to achieve a state change by using a causal reasoning concept. Each of the three must work together to express learned knowledge.

A. Mathematical Formulation for Human Task

Given a set of human task demonstrations $\mathcal{D} = \{D_1, D_2, \cdots, D_n\}$ such as cloth-folding videos, the goal is to learn a joint model (G_{STC}) including Spatial, Temporal, and Causal concepts, that we formulate as

$$\begin{aligned} G^*_{STC} &= \underset{G_{STC}}{\operatorname{argmax}}\ P(G_{STC}|\mathcal{D}) \\ &= P(G_S|\mathcal{D}) \cdot P(G_T|\mathcal{D}) \cdot P(G_C|\mathcal{D}) \\ &\cdot P(R(G_S, G_T, G_C)|\mathcal{D}) \end{aligned} \tag{1}$$

where G_S is the model of spatial concepts, G_T is the model of temporal concepts, G_C is the model of causal concepts, and $R(G_S, G_T, G_C)$ is the relational/conditional model between spatial, temporal, causal concepts.

To implement this formulation, we need to define the concrete representation for each symbol in Eq. 1. Due to the structured and compositional nature of spatial, temporal, and causal concepts, we adopt the hierarchical stochastic grammar model, And-Or graph (AoG) [5], as the base of our model representation which is introduced below. To simplify the learning process, we marginalized the complex STC-AoG (G_{STC}) into the S-AoG (G_S), T-AoG (G_T) and C-AoG (G_C); thus, we can learn the G_S, G_T and G_C separately as the model's initialization, then jointly learn the conditional model between them.

B. And-Or Graph Overview

The And-Or Graph is defined as a 3-tuple $\mathcal{G} = (V, R, P)$, where $V = V^{AND} \cup V^{OR} \cup V^T$ consists of a disjoint set of And-nodes, Or-nodes, and Terminal nodes respectively. R is a set of relations between Or-nodes or subgraphs, each of which represents a generating process from a parent node to its children nodes. $P(r)$ is an expansion probability for each relation.

Figure 1 shows an example of an And-Or graph. An And-node represents the decomposition of a graph into multiple sub-graphs. It is denoted by an opaque circle, and all the out-going edges are opaque lines. An Or-node is a probabilistic switch deciding which of the sub-graphs to accept. It is denoted by an open circle with out-going edges drawn in dashed lines. The Terminal node represents grounded components, often referred to as a dictionary.

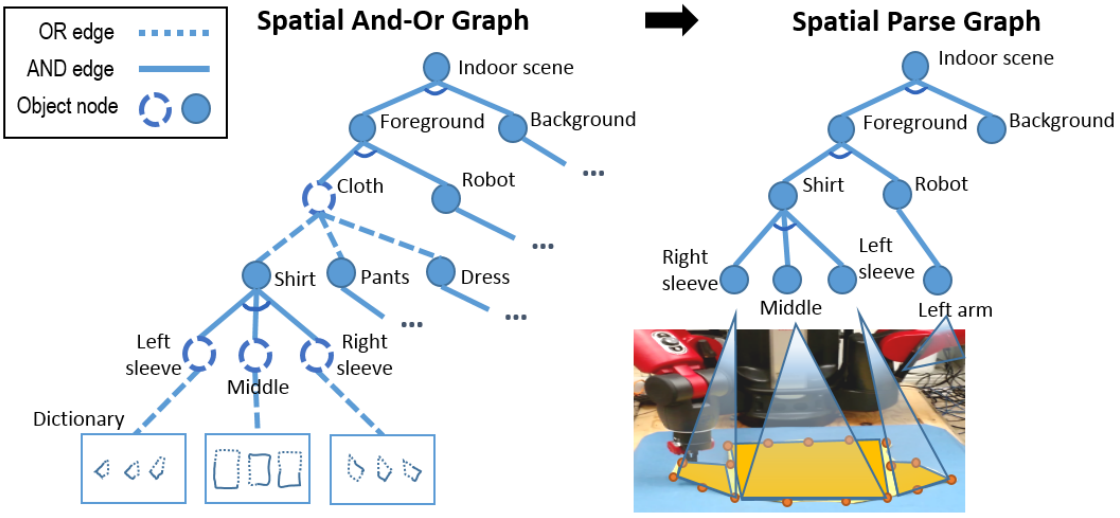


Fig. 1. The Spatial And-Or Graph on the left represents the ongoing perceptual knowledge of the world, i.e. a learned stochastic visual grammar. A specific instance of the And-Or graph is realized in the parse graph on the right.

The nodes are structured into a hierarchical directed acyclic graph (DAG) structure. The AoG is a combination of a Markov tree and Markov random field, where an And-node corresponds to a graphic template model, and an Or-node corresponds to a switch in a Markov tree [17].

Given a set of human demonstrations $\mathcal{D}$, the graph $\mathcal{G}$ is composed of an AoG graph structure $\hat{\mathcal{G}}$ and parameters θ. The nodes and rules/edges in the graph structure aim to maximize the objective function, denoted by the posterior probability:

$$P(\mathcal{G}|\mathcal{D}) = P(\hat{\mathcal{G}}, \theta|\mathcal{D}) \tag{2}$$

$$= P(\hat{\mathcal{G}}|\mathcal{D})P(\theta|\mathcal{D}, \hat{\mathcal{G}}) \tag{3}$$

The first term models the structure of an And-Or graph $\mathcal{G}$ from a human demonstration $\mathcal{D}$. To solve the first term, we manually design the structure of the S-AoG, but we learn the T-AoG and C-AoG structure automatically [4] [25] [15].

The second term models the parameters θ in the graph, given the learned knowledge graph structure. It is reformulated as follows:

$$P(\theta|\mathcal{D}, \hat{\mathcal{G}}) \propto \prod_{D_i \in \mathcal{D}} P(D_i|\theta, \hat{\mathcal{G}}) \tag{4}$$

$$\approx \prod_{D_i \in \mathcal{D}} \max_{pg_i} P(D_i|pg_i, \theta, \hat{\mathcal{G}}) P(pg_i|\theta, \hat{\mathcal{G}}) \tag{5}$$

where pg_i is the parse graph of D_i. A parse graph is an instance of $\mathcal{G}$ where each Or-node decides one of its children. $P(pg_i|\theta, \hat{\mathcal{G}})$ is the prior probability distribution of parse graph pg_i given $\mathcal{G}$. To simply the learning process, we set it as a uniform distribution. Thus,

$$P(\theta|\mathcal{D}, \hat{\mathcal{G}}) \propto \prod_{D_i \in \mathcal{D}} \max_{pg_i} P(D_i|pg_i, \theta, \hat{\mathcal{G}}) \tag{6}$$

And,

$$P(D_i|pg_i, \theta, \hat{\mathcal{G}}) = \prod_{v \in V^{AND}} P(Ch_v|v, \theta_v^{AND}) \tag{7}$$

$$\prod_{v \in V^{OR}} P(Ch_v|v, \theta_v^{OR}) \tag{8}$$

$$\prod_{v \in V^{T}} P(D_i|v) \tag{9}$$

where Ch_v denotes the child of a non-terminal node $v \in V^{AND} \cup V^{OR}$. The probability derivation represents a generating process from a parent node to its child node, and stops at the terminal nodes to generate the sample D_i. The parameters are learned in an iterative process through a Minimax Entropy algorithm explain in more detail later.

C. S-AoG: Spatial Concepts Model

A powerful way to capture perceptual information is through a visual grammar to produce the most probable interpretations of observed images. Therefore, we represent spatial concepts through a stochastic Spatial And-Or Graph (S-AoG) [5]. Nodes in the S-AoG represent visual information of varying levels of abstraction. The deeper a node lies in the graph, the more concrete of a concept it represents. An And-node signifies physical compositionality (i.e. a wheel is a part of a car) whereas an Or-node describes structural variation (i.e. a car is a type of vehicle).

As demonstrated in Figure 1, the root node of the S-AoG encompasses all possible spatial states a robot may

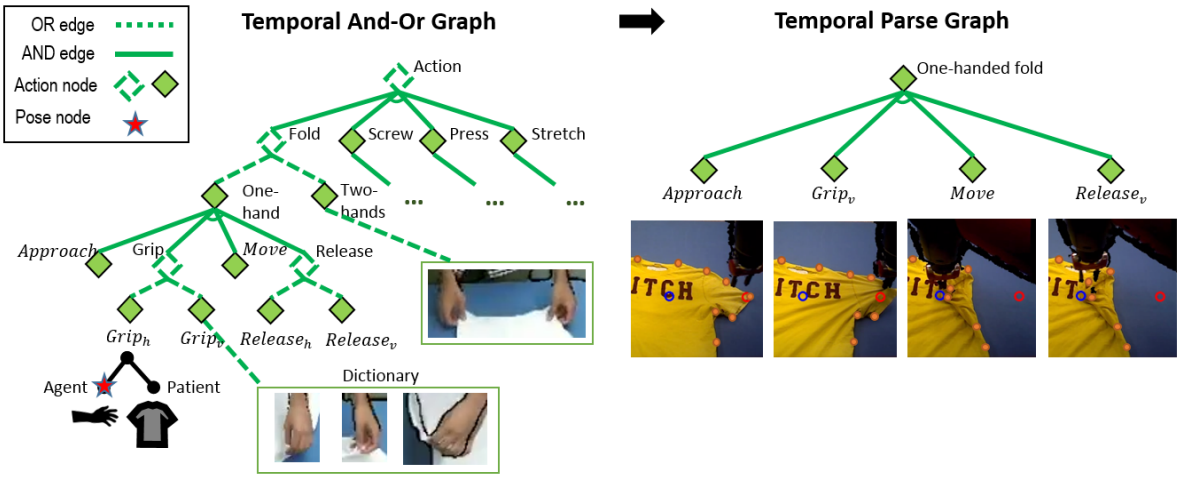


Fig. 2. The Temporal And-Or Graph on the left is a database of all actions currently known in the real world. Each action has an associated agent and patient. The realized parse graph on the right shows a generated sequence of actions directly executable by the robot.

perceive. Here, the "Indoor scene" is decomposed into "Foreground" and "Background," which are then further decomposed. The nodes deeper in the tree represent finer and finer concepts until they end up the terminial nodes consisting of grounded perception units such as the sleeve of t-shirt.

D. T-AoG: Temporal Concepts Model

The action-space of the world is often an assortment of compositional and variational sub-actions. The hierarchical nature of actions leads us to represent actions by a stochastic Temporal And-Or Graph (T-AoG) [4]. And-nodes correspond to a sequence of actions (i.e. close the door, then lock it), whereas Or-nodes correspond to alternate conflicting actions (i.e. close the door, or open the door). The leaf nodes of this graph are atomic action primitives that the robot can immediately perform. Different sequences of atomic actions produce different higher-level actions.

The T-AoG structure is learned automatically using techniques from Si et al. [4] establishing an initial knowledge base of actions. Our T-AoG does not learn new atomic actions, but may learn higher-level actions that are built from these atomic actions. By fixing the set of atomic actions, we ensure the grounding of higher-level actions to alleviate the correspondence problem. Our framework assumes detectors of such atomic action as input.

As shown in Figure 2, the root node of the T-AoG represents all possible actions. As we traverse the tree down, the actions become less and and less abstract, until they can no longer be simplified. Therefore, the robot can unambiguously perform the atomic actions represented by the leaf nodes.

The T-AoG provides us a way to define the structure and sequence of actions, but how an action causes a change in state is incorporated in the causality data structure defined next.

E. C-AoG: Causal Concepts Model

Causality is defined as a fluent change due to a relevant action. We can think of fluents as functions on a situation $x_1(s), x_2(s), ...$, such as the state of a car's engine (on vs. off) or its current speed (5mph, 10mph, etc.). We use the Causal And-Or Graph (C-AoG) to encapsulate causality learned from human demonstration [15], as shown in Figure 3. Each causal node is a fluent change operator, transforming an input fluent to an output fluent by using an action from the T-AoG. As shown in the diagram, there are various ways to reach the same state. Or-nodes capture the various ways a fluent may change from one state to another.

From the point of view of automated planning, fluents are multi-variate observations of a state. The fluents that change due to a relevant action are vital for predicting future actions. If a fluent does not change from a change-inducing action, then it is irrelevant with respect to the action. These time-invariant properties as defined as "attributes" of the node (i.e. color, weight). Additionally, fluents that change due to an inertial action (i.e. actions that are irrelevant to a fluent change) are noted inconsistent.

For example, given an cloth s, let fluent $x_1(s)$ represent high-level abstract information such as the shape of a cloth, whereas if the cloth is a shirt, fluent $x_2(s)$ represents specific keypoints for shirts. The C-AoG structure is learned through an information projection pursuit outlined by Fire et al [15]. The STC-AoG uses

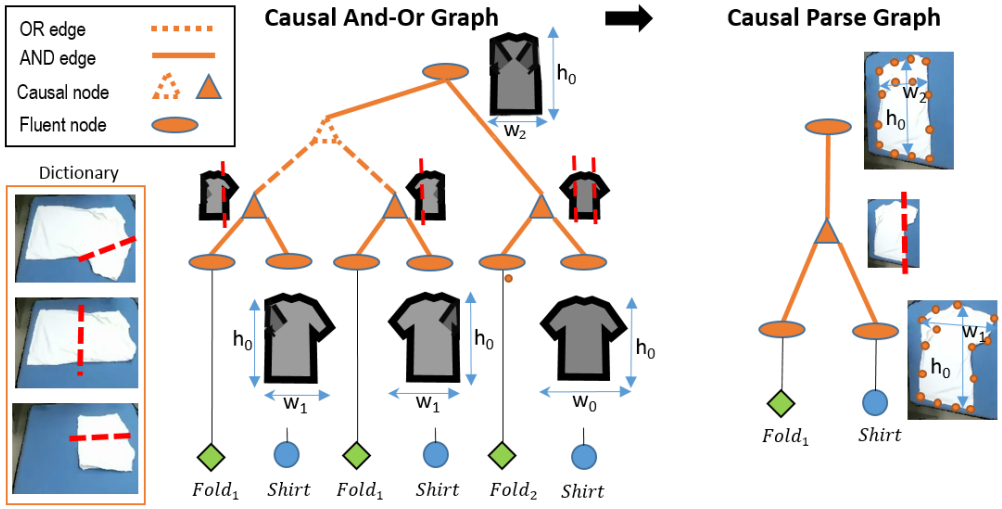


Fig. 3. The Causal And-Or Graph encapsulates the fluent changes per action. The parse graph on the right shows the reasoning system in action.

these relevant fluent changes to plan out tasks.

F. Relational Model between Spatial, Temporal, Causal And-Or Graph

Each of the three And-Or Graphs are unified into a common framework for a complete representation of the world [12]. This explicit knowledge is represented by a hierarchical graphical network specifying a stochastic context sensitive grammar [16], called the the Spatial, Temporal, and Causal And-Or Graph (STC-AoG) [12]. The cloth-folding task in our real-time robot framework is incorporated as described in Figure 4.

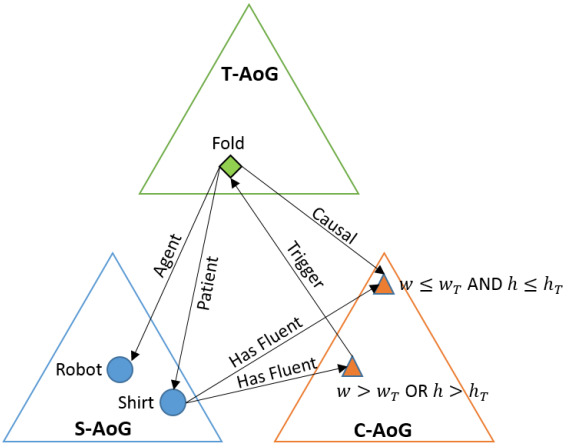


Fig. 4. For illustrative purposes, this diagram shows simple interactions between the spatial, temporal, and causal And-Or graphs. When the width w or height h of the shirt is larger than the target width w_T or height h_T, the C-AoG triggers a fold action in an attempt to reach a smaller folded shirt. The robot then folds the shirt to produce the desired width and height ($w \leq w_t$ AND $h \leq h_T$).

Formally, the fluent functions $\forall j\ x_i(s_j)$ partition the reals $\mathbb{R}$. Two fluents $x_i(s_a)$ and $x_i(s_b)$ are identical if they belong in the same partition. Each spatial or temporal situation s_i may have multiple fluents $(x_1, x_2, ...)$.

$$x(s_i) = \begin{pmatrix} x_1(s_i) \\ x_2(s_i) \\ ... \end{pmatrix} \tag{10}$$

The fluent change between two states s_j and s_k is formally defined as a binary vector:

$$\triangle x(s_j, s_k) = \begin{pmatrix} \triangle x_1(s_j, s_k) \\ \triangle x_2(s_j, s_k) \\ ... \end{pmatrix} \tag{11}$$

$$\triangle x_i(s_j, s_k) = \begin{cases} 0 & \text{if } x_i(s_j) = x_i(s_k) \\ 1 & \text{otherwise} \end{cases}$$

By accumulating human demonstrations of an action, we obtain a set of video clips $Q_a = \{q_1, q_2, ...\}$ for a specific action a, where q_i is a video clip showing action a. The score $w_j(a)$ of an action to make a fluent change is defined as:

$$\forall j\ w_j(a) = P(\triangle x_j = 1 \mid Q) = \frac{\sum_i 1_{\triangle x_j = 1 | q_i}}{||Q||} \tag{12}$$

with the scores normalized by $\sqrt{\sum_j w_j(a)^2}$.

Fluents that represent specific properties, such as keypoints, tend to be heavier weighted than those that are broad high-level concepts, such as shape [18]. The fluents are typically hand-chosen, but we suggest automatically generating various abstractions of fluents by varying the dimensionality of autoencoders. Recent work on spatial semantics [27] can also initialize nodes with a set of useful fluents.

The STC-AoG is not just a knowledge representation system, but also a hierarchical planning graph. Folding a shirt using shirt fluents $x_1(s)$ and $x_2(s)$ has greater affordance than that from using just abstract shape information $x_1(s)$. That way, causal reasoning remains specific to the object, guaranteeing that when folding a shirt, there is less preference to use knowledge about how to fold pants if knowledge about how to fold shirts already exists. We define the affordance of transferring from state s_i to s_j using action a by $\textbf{aff}(a, s_i, s_j) = w(a)^T \triangle x(s_i, s_j)$, suggesting that the automated planning and reasoning should only be based on the relevant features.

Unifying the three sub-graphs produces a closed-loop framework for robots learning from demonstrations. Moreover, graphs can store relationships in an intuitive and highly regular structure, allowing for algorithms that rely on simple graph manipulations. The real world is encoded through perception into the S-AoG to form a physical belief state of the world. The learning algorithm constructs a C-AoG to understand actions from human demonstrations. And lastly, inference combines the reasoning from the C-AoG and the actuators from the T-AoG to physically perform the task. The energy of the joint parse graph [12] combines the energy terms of each:

$$E_{STC}(pg) = E_S(pg) + E_T(pg) + E_C(pg) + \sum_{r \in R^*_{pg}} E_R(r) \tag{13}$$

We use generative learning by the Minimax Entropy Principle [20] to learn the probability distribution of STC parse graphs $P(pg)$. Doing so assumes that the sample mean of statistics $\phi_j(pg)$ should approach the true expectation s_j from observations. The parameters are solved by minimizing the Kullback-Leibler divergence between the observed distribution and the candidate $KL(f||p) = E_f[\log f(pg)] - E_f[\log p(pg)]$. This simplifies to a maximum likelihood estimate, formulated by

$$p^* = \underset{p \in \Omega}{\text{argmax}}\ E_f[\log p(pg)] = \underset{p \in \Omega}{\text{argmax}} \sum_{i=1}^{n} \log p(pg_i) + \epsilon \tag{14}$$

Iteratively, we choose the statistics $F = \{\phi_1, \phi_2, ...\}$ that minimize the entropy of the model, and the parameters β that yield maximum entropy.

$$p^* = \underset{F}{\text{argmin}}\{\max_{\beta}\ \text{entropy}(p(pg; \theta)\} \tag{15}$$

Effectively, the robot "daydreams" possible probability distributions of parse graphs to converge with observations. During inference, it samples a parse graph to perform the action.

G. Learning Motor Control

The STC-AoG expresses explicit knowledge in a graphical structure easily understandable by humans, acting as a gateway for communication. However, the STC-AoG only defines discrete salient spatial, temporal, and causal concepts. The interpolation of how an individual action is performed requires a specification of the fine motor skills involved as well as an assignment of probability distribution parameters.

The explicit knowledge captured by a causal node represents a conformant plan learned by human demonstrations. The information stored in the STC-AoG only provides results from discrete time-steps, $t \in \mathbb{N}$. Its state-action table represents fluent changes by $x^{t+1}(s) = f(x^t(s), x^t(a))$. To shift paradigms from explicit to implicit knowledge, we relax the assumption of null runtime observability, and use a finer distinction in time, $x^{t+\delta t}(s) = f(x^t(s), x^t(a))$. By learning this continuous function f, the robot system is capable of verifying, correcting, and inferring causal relations to adapt to dynamic environments.

We make two assumptions to simplify the learning of f. First, we restrict the range of spatial and temporal changes to adhere to spatiotemporal continuity, rendering sudden changes impossible. Second, we use a physical simulator based on perception encoded by the STC parse graph (STC-pg) to compare with reality at rapid time intervals. When a discrepancy is detected, we point fault at the robot's actions. The feedback learning system uses a simplified optimization process inspired by Atkeson et al [22] to update the control mechanics. Adjusting the parameters of the simulator to adhere to reality also reveals useful knowledge, but it is out of scope for this study.

H. Inference

Since the STC-AoG model is generatively learned, we infer a parse graph through a simple sampling process. As seen in Figure 5, the procedurally generated parse graph lays out a conformant action plan for the robot. It then creates a simulation of the action by converting the STC-pg into a motion plan and spatial objects into 3D meshes from point cloud.

The simulation plan is matched with reality at small interval steps to verify that the robot is at its corresponding simulated state. In case of substantial mismatch between expected and actual states, the robot understands the action did not complete, and that a new action plan must be generated based on the latest perception input. Concretely, the sampling procedure is encapsulated by the algorithm in Figure 6.

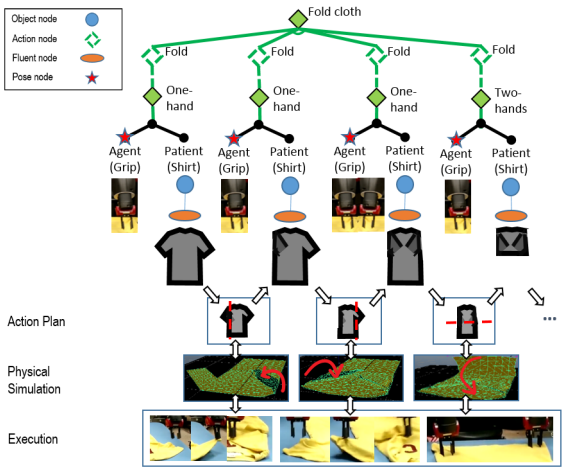


Fig. 5. The inference engine samples a parse graph to create a conformant action plan. There is feedback between the plan, its simulation, and the corresponding perceived execution.

1: **while** camera is producing image I **do**
2: $pg_S^t \leftarrow \text{Interpret}(G_S, I_t)$
3: $pg_T^t \leftarrow \text{Sample}(G_{STC}, pg_S^t)$
4: $pg_C^t \leftarrow \text{Sample}(G_{STC}, pg_S^t, pg_T^t)$
5: $pg_{STC} \leftarrow \text{Merge}(pg_S^t, pg_T^t, pg_C^t)$
6: PerformWithFeedback(pg_{STC})
7: **end while**

Fig. 6. The robot inference algorithm performs tasks on a learned STC-AoG. It interprets the sensory input as spatial, temporal, and causal parse graphs, which are merged to formed a joint representation that is sampled and acted on.

IV. EXPERIMENTS

We conduct our experiments on a cloth-folding task. The S-AoG models the physical status of the cloth, table, robot, human, and various decompositions of each. The T-AoG consists of three atomic actions to span the action-space for this simple task: *MoveArm(a)*, *Grab*, and *Release*. A *Fold* action in the T-AoG is a higher-level And-node consisting of four children: *MoveArm(a)*, *Grab*, *MoveArm(b)*, and *Release*, with the corresponding textual representation: $Fold(a, b) = MoveArm(a); Grab; MoveArm(b); Release$. And consequently, a specific instance of folding is a series of *Fold* actions: $FoldStyle1 = Fold(a, b); Fold(c, d); ...; Fold(y, z)$. Lastly, the C-AoG nodes describe how to fold a shirt from one state to another, learned through human demonstrations.

We use Baxter, a two-armed industrial robot to perform our cloth-folding task. Each arm consists of 7 degrees of freedom that are adjusted through inverse kinematics relative to the robot's frame of reference. The robot's primary perception sensor is an Asus PrimeSense camera that provides an aligned RGB-D (Red, Green, Blue, and Depth) point cloud in real-time. In order to use localization results from perception, we compute the affine transformation matrix from the camera coordinate system to that of the robot. All components interact together through the Robot Operating System (ROS).

The STC-AoG is stored in the platform-independent Graphviz DOT language, and used by our platform written in C++. The hand-designed perception logic combines off-the-shelf graph-based [24] and foreground/background [23] segmentation to localize a cloth per frame. On top of that, we train a shirt detector model using a Support Vector Machine to facilitate narrowing down the search for an optimal S-AoG parse graph. Each cloth node has a fluent x_1 describing the low-level shape. If a cloth is a shirt, we represent the structure of its keypoints as another fluent x_2. We simplify learning the probability distribution of parse graphs by limiting the number of statistics to $F = \{\phi_1\}$, where ϕ_1 is the affordance cost of the action sequence in a STC-pg.

Performance on a task is measured by the percent of successful actions throughout the task. The overall performance is the average of all task performances over multiple trials. An action is successful if performing the action satisfies the pre- and post-conditions of the causal relationship used.

A. Experiment Settings

In the first set of experiments, we measure the performance of representing learned knowledge from human demonstrations. After watching human demonstrations, the robot generates an action plan step by step. The human performs the action suggested by the robot, and at each step, the human qualitatively verifies whether the robot's action was indeed the intended action as per the demonstration. If verification fails in either case, then the action is marked unsuccessful, and otherwise it is marked successful. This performance score on learning will set the baseline for the next set of experiments.

In the second series of experiments, we measure the quality of grounding the learned knowledge to the robot's actions. This time we let the robot, instead of the human, perform the actions. We compare the performance of the robot folding clothes with the results from the first set of experiments to evaluate the success of grounding physical actions to see how well they match that of a human. The expected performance should be less than the ground truth established from the previous experiment.

In the third series of experiments, we measure the improvements from a feedback system compared to no feedback. We expect that the performance score calculated through this step should be higher than that from the previous experiment, but lower than the ground truth.

Finally, we are also curious how much we can stretch the generalizability of a learned task. After demonstrating how to fold a t-shirt, we ask the robot to infer how to fold different articles of clothing, such as full-sleeve shirts, towels, and pants. The criteria for generalizability of knowledge will follow the similar performance procedure as in the previous experiments.

B. RESULTS

On 10 trials per four sets of different t-shirt folding demonstrations D_1, D_2, D_3, D_4, we measure the average performance of using our system to learn knowledge, ground robot actions, and control feedback.

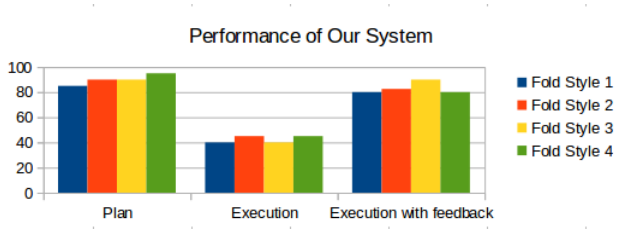


Fig. 7. Our learning system successfully understood the various folding techniques. It had some difficulty executing the task using simply a conformant plan, but with added feedback the execution was highly successful.

As seen in Figure 7, our knowledge representation system was able to characterize the cloth-folding task enough to faithfully communicate with a human, producing a learned representation with an average performance of 90%. This sets the upper bound for the next two inference experiments. As anticipated, our framework was able to ground the actions with a performance of 42.5%. The low score indicates that although the robot knows what to do, there is still a discrepency between the human's action and that generated by the STC-AoG. By adding feedback correction of comparing perception to physical simulation, the performance leaped to 83.125%, also matching our expectation.

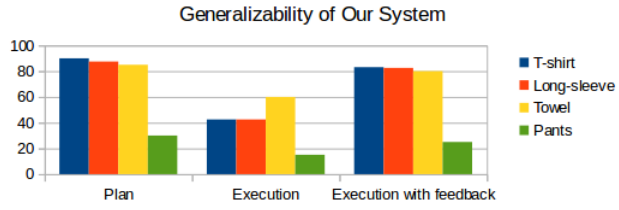


Fig. 8. Our knowledge framework correctly understood how to generalize a t-shirt folding instruction to long-sleeve shirts and towels; however, it expectedly had difficulty extrapolating its knowledge to fold pants.

The performance of generalizability was measured after training the robot on only t-shirt folding videos. The results are visualized in Figure 8. For example, since a full-sleeve shirt may have the same width and height fluents as that of a t-shirt, the inference plan for folding a full-sleeve shirt performed very well. Moreover, the robot was able to generate reasonable action plans to fold a towel it has never seen, since a t-shirt with both its sleeves folded resembles the same rectangular shape of a towel. However, generating a reasonable inference result for folding pants was less successful due to the natural lack of knowledge transferred between a shirt folding and pant folding task. Figure 9 shows a few qualitative results of successful folding plans and executions.

Fig. 9. Some qualitative results on the robot execution after learning from human demonstrations.

V. DISCUSSION AND FUTURE WORK

The experiments show preliminary support for the expressive power of the robot learning and execution framework laid out in this paper. While we focus heavily in the cloth-folding domain, the framework may be used for training any goal-oriented task. In future work, we wish to continue improving the robustness of each spatial, temporal, and causal And-Or graph to optimize for speed and accuracy.

The STC-AoG acts as a language to ground knowledge and reasoning into robot actions. Since the knowledge representation and robot action planning systems share the same And-Or graph data structure, the graph acts as a programming language for the robot, and self-updating the graph is an act of metaprogramming.

Due to the hierarchical nature of the STC-AoG, the higher level nodes are readily articulated and understandable by humans. are currently working on incorporating natural language statements, commands, and questions to more easily allow humans to manipulate the graph. To scale up the graph for life-long learning, we are investigating other practical storage solutions, including graph-based databases such as Neo4j [30]. Since the graph is sufficient to transfer knowledge, we can upload different skills to a cloud platform and share knowledge between different robots.

Limits in physical reachability and dexterity of the robot arms played a crucial difficulty in mapping action plans to motor control execution. If a grip location was unreachable, the conformant plan would fail to execute the action at all. Fortunately, by introducing the feedback control system, we were able to at least extend the reach as far as possible to grip a reasonable point.

Lastly, the performance of the causal learning system relies on successfully detecting fluent changes. This requires adjusting thresholds for fluent-change detectors until the results seem just right. We solved this problem by offline supervised learning for our chosen fluents, but we set aside the problem of learning these threshold parameters online to future work.

VI. CONCLUSIONS

The stochastic graph-based framework is capable of representing task-oriented knowledge for tractable inference and generalizability. It successfully unified theoretical foundations of And-Or perception grammars to a practical robotics platform. The experimental results support our claims for grounding learned knowledge to execute tasks accurately. We also express the generalizability of our framework by extrapolating from human demonstrations of folding a t-shirt to other articles of clothing. And lastly, our novel framework can make use of perceived discrepancies between high-level action plans and low-level motor control to verify and correct actions.

ACKNOWLEDGMENT

The authors would like to thank the support of DARPA SIMPLEX project N66001-15-C-4035 and DARPA MSEE project FA 8650-11-1-7149. In addition, we would like to thank SRI International and OSRF for their support.

REFERENCES

[1] E. Theodorou, J. Buchli, and S. Schaal, "Reinforcement learning of motor skills in high dimensions: A path integral approach," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2397–2403.

[2] D. Kulić, C. Ott, D. Lee, J. Ishikawa, and Y. Nakamura, "Incremental learning of full body motion primitives and their sequencing through human motion observation," *The International Journal of Robotics Research*, p. 0278364911426178, 2011.

[3] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 37, no. 2, pp. 286–298, 2007.

[4] Z. Si, M. Pei, B. Yao, and S.-C. Zhu, "Unsupervised learning of event and-or grammar and semantics from video," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 41–48.

[5] S.-C. Zhu and D. Mumford, "A stochastic grammar of images," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2006.

[6] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos, "Robot learning manipulation action plans by unconstrained videos from the world wide web," in *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[7] N. Shukla, C. Xiong, and S.-C. Zhu, "A unified framework for human-robot knowledge transfer," in *AAAI'15 Fall Symposium on AI for Human-Robot Interaction (AI-HRI 2015)*, 2015.

[8] C. Liu, J. Y. Chai, N. Shukla, and S.-C. Zhu, "Task learning through visual demonstration and situated dialogue," in *AAAI'16 Workshop on Symbiotic Cognitive Systems*, 2016.

[9] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale image classification," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 177–186.

[10] A. L. Blum and M. L. Furst, "Fast planning through planning graph analysis," *Artificial Intelligence*, vol. 90, no. 1, pp. 281–300, 1997.

[11] S. Miller, J. van den Berg, M. Fritz, T. Darrell, K. Goldberg, and P. Abbeel, "A geometric approach to robotic laundry folding," *International Journal of Robotics Research (IJRR)*, vol. 31, no. 2, pp. 249–267, 2012.

[12] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *MultiMedia, IEEE*, vol. 21, no. 2, pp. 42–70, 2014.

[13] J.-W. Ha, K.-M. Kim, and B.-T. Zhang, "Automated construction of visual-linguistic knowledge via concept learning from cartoon videos," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015), Austin*, 2015.

[14] G. W. Strong and Z. W. Pylyshyn, "Computation and cognition: Toward a foundation for cognitive science. cambridge, massachusetts: The mit press, 1984, 320 pp." *Behavioral Science*, vol. 31, no. 4, pp. 286–289, 1986.

[15] A. S. Fire and S. Zhu, "Learning perceptual causality from video," in *Learning Rich Representations from Low-Level Sensors, Papers from the 2013 AAAI Workshop, Bellevue, Washington, USA, July 15, 2013*, 2013.

[16] J. Rekers and A. Schrr, "A parsing algorithm for context-sensitive graph grammars," Tech. Rep., 1995.

[17] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 943–950.

[18] C. Chao, M. Cakmak, and A. Thomaz, "Towards grounding concepts for transfer in goal learning from demonstration," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, vol. 2, Aug 2011, pp. 1–6.

[19] E. T. Mueller, in *Commonsense Reasoning*, E. T. Mueller, Ed. Morgan Kaufmann, 2006.

[20] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling." *Neural Computation*, vol. 9, no. 8, pp. 1627–1660, 1997.

[21] R. P. N. Rao, A. P. Shon, and A. N. Meltzoff, "A bayesian model of imitation in infants and robots," in *In Imitation and Social Learning in Robots, Humans, and Animals*. Cambridge University Press, 2004, pp. 217–247.

[22] C. Atkeson and S. Schaal, "Learning tasks from a single demonstration," in *Robotics and Automation, 1997. Proceedings., 1997 IEEE International Conference on*, vol. 2, Apr 1997, pp. 1706–1712 vol.2.

[23] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut -interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH)*, August 2004.

[24] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sept. 2004.

[25] K. Tu, M. Pavlovskaia, and S.-C. Zhu, "Unsupervised structure learning of stochastic and-or grammars," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 1322–1330.

[26] Y. Yamakawa, A. Namiki, and M. Ishikawa, "Motion planning for dynamic folding of a cloth with two high-speed robot hands and two high-speed sliders." in *ICRA*. IEEE, 2011, pp. 5486–5491.

[27] K. Zampogiannis, Y. Yang, C. Fermller, and Y. Aloimonos, "Learning the spatial semantics of manipulation actions through preposition grounding." in *ICRA*. IEEE, 2015, pp. 1389–1396.

[28] A. Doumanoglou, A. Kargakos, T. Kim, and S. Malassiotis, "Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning," in *2014 IEEE International Conference on Robotics and Automation,*

ICRA 2014, Hong Kong, China, May 31 - June 7, 2014, 2014, pp. 987–993.

[29] P. C. Wang, S. Miller, M. Fritz, T. Darrell, and P. Abbeel, "Perception for the manipulation of socks." in *IROS*, 2011, pp. 4877–4884.

[30] Neo4j, *Neo4j - The Worlds Leading Graph Database*, Std., 2012. [Online]. Available: http://neo4j.org/

Chapter 3: Task Learning through Visual Demonstration and Situated Dialogue

Abstract

To enable effective collaborations between humans and cognitive robots, it is important for robots to continuously acquire task knowledge from human partners. To address this issue, we are currently developing a framework that supports task learning through visual demonstration and natural language dialogue. One core component of this framework is the integration of language and vision that is driven by dialogue for task knowledge learning. This paper describes our on-going effort, particularly, grounded task learning through joint processing of video and dialogue using And-Or-Graphs (AOG).

Introduction

As a new generation of social robots emerges into our daily life, techniques that enable robots to learn task-specific knowledge from human teachers have become increasingly important. In contrast to previous approaches based on Learning from Demonstration (Chernova and Thomaz 2014) and Learning by Instruction (She et al. 2014), we are currently developing a framework that enables task learning through simultaneous visual demonstration and situated dialogue. Supported by our framework, robots can acquire and learn grounded task representations by watching humans perform the task and by communicating with humans through dialogue. The long-term goal is to enable intelligent robots that learn from and collaborate with human partners in a life-long circumstance.

A key element in our framework is And-Or-Graph (AOG) (Tu et al. 2014; Xiong et al. 2016), which embodies the expressiveness of context sensitive grammars and probabilistic reasoning of graphical models. We use AOG to build a rich representation (i.e., STC-AOG) of the Spatial, Temporal, and Causal knowledge about the real world and the task. In addition, we are also designing an AOG-based schema (i.e., CI-AOG) to model and interpret the communicative intents between an agent and its human partner. These expressive and deep representations then allow a robot and a human to efficiently and effectively establish and increment their common ground (Clark 1996) in learning real-world tasks.

This paper provides an overview of the AOG-based framework and uses an example to illustrate our ongoing work on joint task learning from visual demonstration and situated dialogue.

Representations

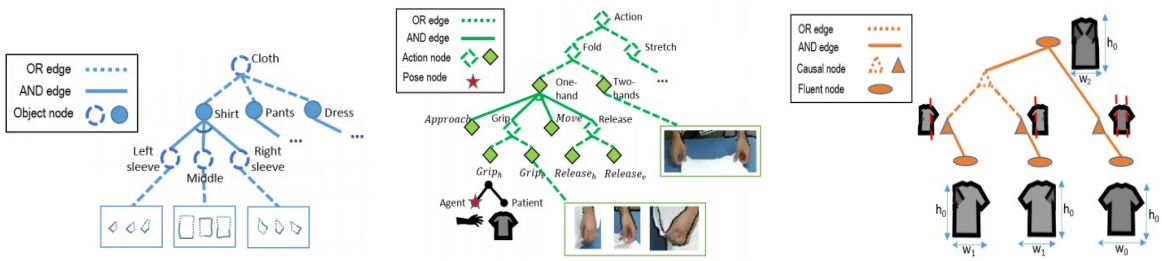
STC-AOG

An *And-Or-Graph* (*AOG*) (Tu et al. 2014) is an extension of a constituency grammar used in Natural Language Processing. It is often visualized as a tree structure consisting of two types of nodes, i.e., *And-node* and *Or-node*. An And-node represents the configuration of a set of sub-entities to form a composite entity; An Or-node represents the set of alternative compositional configurations of an entity. Using this general representation, three important types of task knowledge can be modeled:

- Spatial And-Or Graph (S-AOG) models the spatial decompositions of objects and scenes.
- Temporal And-Or Graph (T-AOG) models the temporal decompositions of events to sub-events and atomic actions.
- Causal And-Or Graph (C-AOG) models the causal decompositions of events and fluent changes.

Figure 1 illustrates an example of the S-/T-/C- AOG representation for cloth-folding tasks, which captures the spatial, temporal, and causal knowledge of the domain. Robots can then utilize this rich knowledge representation to understand, communicate, and perform task-oriented actions. Based on this knowledge representation framework, Xiong et al. (2016) has developed a statistical learning mechanism that automatically learns the parameters (e.g., the branching probabilities of Or-Nodes) of S-/T-/C-AOGs from a set of human demonstration videos. Furthermore, methods for learning the structures of different types of AOG have also been studied in previous work (e.g., Pei et al. 2013; Fire and Zhu 2013).

The basic idea of learning AOG-based task knowledge is to treat each demonstration as a specific instance, or a so-called "parse graph", which is generated by selecting one of the alternative configurations at each



(a) S-AOG (b) T-AOG (c) C-AOG

Figure 1: An example of the S-/T-/C- AOG for a cloth-folding domain.

Or-node of an AOG model (see Tu et al. (2014) for details). Given a series of demonstrations represented as parse graphs, the structures and parameters of the underlying AOG model then can be learned using statistical learning techniques.

CI-AOG

Since AOG in essence can be viewed as a stochastic grammar machinery, and has been shown powerful in parsing the hierarchical structure of goal-driven events (Pei et al. 2013), we propose to use the same mechanism for analyzing the intentional structure of knowledge transferring dialogues.

For this purpose, we first construct an AOG, which we call the "Communicative Intent" AOG (CI-AOG) here, to describe how the intentional structure of such dialogues could possibly unfold. Our CI-AOG is similar to the T-AOG or "event grammar" as we illustrated earlier, where an Or-node captures different possibilities and an And-node captures sequential events, and the terminal nodes represent the basic actions (i.e., dialogue acts) that one can perform in a dialogue.

To illustrated the idea, we have manually crafted a (partial) CI-AOG that can be used to analyze the intentional structure of a task teaching dialogue as shown in Figure 2. We composed this CI-AOG based on "situated learning" literature (Lave and Wenger 1991; Herrington and Oliver 1995) to model how the teacher's and the learner's intents interact in a mixed-initiative dialogue. For example, we capture in this CI-AOG the common intentions in situated learning, such as *articulation* (the leaner articulates what is being understood regarding the current situation), *reflection* (the learner reflects what has been learned), and *assessment* (the teacher provides feedback to the learner's reflections or articulations).

Furthermore, the CI-AOG is also used to capture the unique characteristics of dialogue, including turn-taking, initiatives, and collaborative dynamics (Clark 1996; Clark and Schaefer 1989). To capture the turn-taking dynamics in dialogue, each node in CI-AOG is assigned a role (i.e., who the speaker is). This is illustrated in Figure 2 by assigning different colors to the nodes (i.e., orange nodes represent the learner and blue nodes represent the teacher). Therefore, an And-node in CI-AOG not only represents the temporal order of its

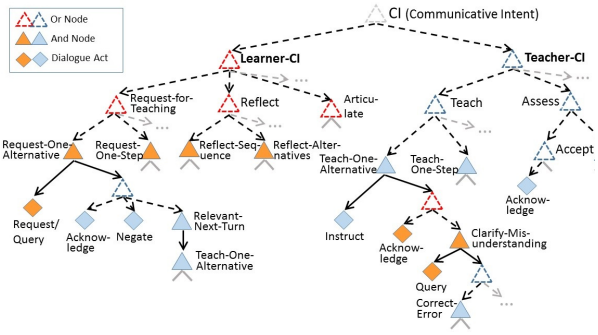


Figure 2: An example of Communicative Intent AOG (CI-AOG).

children nodes, but also captures who takes the initiative of the sub-dialogue and how the turn-taking switches between the learner and the teacher.

The expressiveness of the AOG language also allows us to capture the collaborative dynamics studied in the discourse analysis literature (e.g., Clark and Schaefer (1989)). For example, as illustrated in the left bottom part of Figure 2, after the learner requests the teacher for teaching an alternative way of doing a task (i.e., the *Request-One-Alternative* node), the teacher should respond an explicit acknowledgement, or a negation, or directly teach an alternative without explicit acknowledging (the "relevant-next-turn" behavior).

Suppose a CI-AOG has already been constructed, it then can be used for "parsing" the underlying intentional structure of an ongoing dialogue. This is similar to previous work (Pei et al. 2013) that used a Top-Down parsing algorithm to analyze the hierarchical structure of goal-driven events from an observed sequence of atomic actions. Figure 3 further illustrates a parse graph on the underlying intentional structure of the following example dialogue.

Example dialogue of a robot learning to fold a t-shirt:

R1: Could you teach me how to fold the t-shirt?
H1: Sure.
H2: First, you fold the right sleeve towards the middle of the t-shirt.
R2: I saw you approached the right sleeve, grasped there, then moved to a position, and released your hand there.
R3: Is that position the middle of the t-shirt?

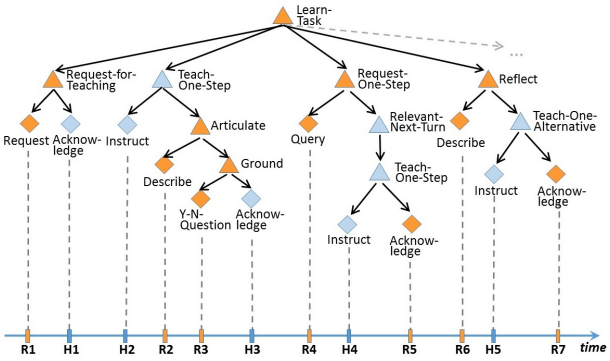


Figure 3: The CI parse graph for the given example dialogue.

H3: Yes, that position is in the middle of the t-shirt.
R4: OK, what is the next step?
H4: Next, you fold the left sleeve to the middle.
R5: OK.
R6: This is what I have learned: first fold the right sleeve to the middle, and then fold the left sleeve to the middle.
H5: You can also fold the left sleeve first, and then the right sleeve.
R7: I see.
R8: What is next?
. . . . . .

As illustrated in Figure 3, the overall intent of this dialogue is for the robot to learn a sequential task. It is initiated by the robot's request for teaching (R1), followed by the human's explicit acknowledgement (H1). The following sub-dialogue is then led by the human's intent of teaching the robot the first step with an instruction (H2). Following that, the robot articulates what it understands about the current situation (R2), and tries to map the unknown concept "middle" to a physical position in the visual context (the question asked in R3, with an intent of what we call "*ground*"). The human's positive response (H3) confirms the robot's understanding, and also closes the subroutine of teaching the first step. The dialogue routine then rolls back to a higher-level of the intent hierarchy, where the robot moves on with its intent of learning the next step (R4). In R6, after two consecutive steps have been learned, the robot issues a reflection on what has been learned so far, which triggers human's following intent to teach an alternative order (H5).

Now we have introduced different types of AOG as the fundamental representations of the physical world, task knowledge, and dialogue dynamics. Next we turn our focus to discussing how we utilize these representations to build learning agents under a unified framework.

Learning from Situated Dialogue

Natural language and dialogue can play an important role in learning task knowledge from a human. Language provides a key source of information to gear the learned knowledge towards how humans conceptualize and communicate about situations and tasks. Such "human-oriented" knowledge is very necessary for facilitating human-robot communication and collaboration (for example, Lemon, Gruenstein, and Peters (2002)).

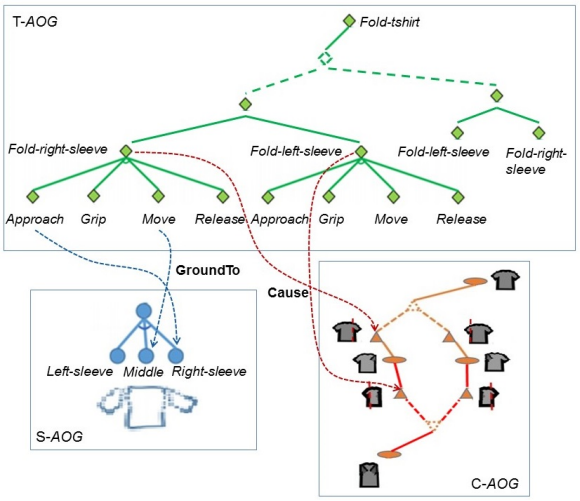


Figure 4: The STC-AOG representation of task knowledge that can be learned from the previous example dialogue of learning to fold a t-shirt. Note that the S-/T-/C- components are not independent from each other. The interplay between them provides an integrated representation of the task knowledge.

Furthermore, dialogue provides an expedited way to learn task knowledge. This can be demonstrated by our earlier example of learning how to fold a t-shirt. After the robot reflected (in R6) the just learned two steps (i.e., *fold-right-sleeve* and *fold-left-sleeve*), the human further taught that the order of the two steps could be switched and it would result into the same status of performing the task (H5). With our AOG-based representation, the robot can add this new knowledge by directly modifying the high-level structure of the STC-AOG (i.e., create new temporal and causal Or-Nodes to represent this alternative sequence of actions and fluent changes). Using language makes it much easier to communicate such high-level knowledge (Figure 4 illustrates the STC-AOG representation that can be learned thereafter).

We thus propose an AOG-based framework to enable robot learning task knowledge from natural language and visual demonstration simultaneously. Supported by this framework, the robot can also proactively engage in human's teaching through dialogue, and gradually accumulate and refine its knowledge. One key advantage of our proposed framework is to provide a unified view of modeling the joint and dynamic task learning process. Besides, since we use AOG as a common representation basis, different components of our model can be stored and accessed using the same format (e.g., graph database), and be processed by the same set of algorithms. It thus can greatly ease the burden of building

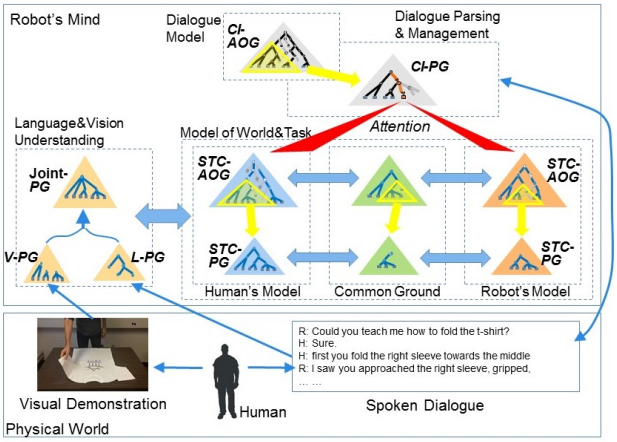


Figure 5: Illustration of our AOG-based framework for supporting robot learning from situated dialogue.

complex AI agents.

Figure 5 illustrates the basic ideas of our task learning system. It mainly consists of three tightly connected components that are all based on AOG representation and processing:

- *Language and Vision Understanding* processes the visual context into a "Vision Parse Graph" (*V-PG*) and the linguistic context into a "Language Parse Graph" (*L-PG*), and fuses them together into a "Joint Parse Graph" (*Joint-PG*) for a deep and accurate understanding of the current situation. A previous work (Tu et al. 2014) has employed the same AOG-based representations for joint text and video parsing in the question-answering domain. The processing in our component here resembles that work. However the linguistic content of a dialogue could require more sophisticated approaches than those for handling monologues, and our goal is to learn generalizable task knowledge rather than just understand one situation.
- *World and Task Model* manages the representation and acquisition of knowledge of the physical world and tasks. As introduced earlier, we use STC-AOG to represent general knowledge about the world and the tasks, while a specific situation (i.e., a Joint Parse Graph) is represented as an instantiation (or subgraph) of the STC-AOG. Motivated by the Common Ground theory (Clark 1996), our agent maintains three copies of models. One is the human's model of the world and knowledge, which is inferred from the joint parsing of language and vision. One is the agent's own model, and the third one is their shared/matched understanding of the situation and knowledge of the task (i.e., their *common ground*). In future work, we will further extend these models towards modeling the "Theory of Mind" in human-robot collaboration.
- *Dialogue Modeling and Management* uses CI-AOG to model and analyze the intentional structure of the task learning dialogue, and to facilitate the agent's decision making in knowledge acquisition and dialogue engagement. Our design of the situated dialogue agent also resembles the classical theory on discourse modeling (Grosz and Sidner 1986). I.e., the *intentional structure* is captured by a CI- Parse Graph (CI-PG) in our dialogue management component. The *linguistic structure* in our case has been extended to the joint model of the linguistic and visual contexts (captured as STC- Parse Graphs), and the shared knowledge (captured as STC-AOG). The *attentional state* is captured by linking each node in the CI-PG to a specific node/edge in the situation or knowledge representation graphs.

As the dialogue and demonstration unfold, the agent dynamically updates its intent, situation, and knowledge graphs. Each component can utilize the information from others through the interconnections between their graph representations. Based on this unified framework, sophisticated learning agents can become easier to be designed and built.

Conclusion

This paper provides a brief overview of our on-going investigation on integrating language, vision, and situated dialogue for robot tasking learning based on And-Or-Graphs (AOG). In particular, through an example, it demonstrates how language and dialogue can be used to augment visual demonstration by incorporating higher-level knowledge. Here we use cloth-folding as an example, but the same framework can be extended to other types of task learning. We are currently in the process of implementing the end-to-end system and plan to collect realistic data to evaluate our approach.

Acknowledgment

This work was supported by a DARPA SIMPLEX grant N66001-15-C-4035.

References

Chernova, S., and Thomaz, A. L. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8(3):1–121.

Clark, H. H., and Schaefer, E. F. 1989. Contributing to discourse. *Cognitive science* 13(2):259–294.

Clark, H. H. 1996. *Using language*. Cambridge university press.

Fire, A. S., and Zhu, S.-C. 2013. Learning perceptual causality from video. In *AAAI Workshop: Learning Rich Representations from Low-Level Sensors*.

Grosz, B. J., and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* 12(3):175–204.

Herrington, J., and Oliver, R. 1995. Critical characteristics of situated learning: Implications for the instructional design of multimedia.

Lave, J., and Wenger, E. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Lemon, O.; Gruenstein, A.; and Peters, S. 2002. Collaborative activities and multi-tasking in dialogue systems: Towards natural dialogue with robots. *TAL. Traitement automatique des langues* 43(2):131–154.

Pei, M.; Si, Z.; Yao, B. Z.; and Zhu, S.-C. 2013. Learning and parsing video events with goal and intent prediction. *Computer Vision and Image Understanding* 117(10):1369–1383.

She, L.; Yang, S.; Cheng, Y.; Jia, Y.; Chai, J. Y.; and Xi, N. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 89.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S.-C. 2014. Joint video and text parsing for understanding events and answering queries. *MultiMedia, IEEE* 21(2):42–70.

Xiong, C.; Shukla, N.; Xiong, W.; and Zhu, S.-C. Robot learning with a spatial, temporal, and causal and-or graph. Submitted to ICRA 2016.